DETECTING NOVELTIES ON PLANETARY SURFACES WITH AUTOENCODERS

Virtual Conference 19–23 October 2020

Braden Stefanuk[1], Alexis Pascual[2], Krzysztof Skonieczny[1], Kenneth McIsaac[2], Kaizad Raimalwala[3], Evan Smal[3], Michele Faragalli[3]

[1]*Concordia University, 1455 De Maisonneuve Blvd. W., Montreal, Canada, Email: bstefanuk@gmail.com*
[2]*Western University, 1151 Richmond St, London, Canada, Email: apascua2@uwo.ca*
[3]*Mission Control Space Services, 162 Elm St., Ottawa, Canada, Email: kaizad@missioncontrolspaceservices.com*

ABSTRACT

In the domain of planetary science, novelty detection is quickly gaining attention because of the operational solutions it offers, including annotated data products and downlink prioritization. When detecting novelties in images, autoencoders have shown to have value, both in their predictive properties and their visualizations. In this study, a processing pipeline that supports rapid autoencoder prototyping for novelty detection is presented, along with two autoencoder variants. Models are trained on two planetary datasets: The Moon and Mars. Results show that these networks outperform state-of-the-art networks by over 6% when monitoring the area under the receiver operating characteristic curve. Operational viability of the proposed autoencoder networks is discussed.

1 INTRODUCTION

Recent missions to Mars and the Moon have paved the way for future scientific exploration missions to be conducted more autonomously, more precisely, and more robustly. One way this can be achieved is through onboard novelty detection. For example, a vision system capable of detecting planetary surface features of interest, and from these ranking them according to their scientific potential would relieve operational bottlenecks and save precious human time. Such a system could intelligently prioritize which content is downlinked, allowing mission operators to spend less time with manual selection. It could supplement the process of scientific analysis with automated predictions, and perhaps most importantly, it could enable data collection to continue during communication dropouts.

Previous work in the area of image-based novelty detection has demonstrated that convolutional autoencoders (CAEs) can detect Martian surface features on par with other methods such as Reed-Xiaoli detectors and Generative Adversarial Networks (GANs) [1]. One benefit of autoencoders over these other methods is the interpretability of their results. Since the training objective of an autoencoder is to reproduce the input image from a lower dimensional (latent) representation, once an autoencoder has been trained, error maps can be created to visualize how the reconstructed outputs and the original inputs differ. As can be seen in Fig. 1, these error maps supply spatial and spectral information about the location and magnitude of novel features.

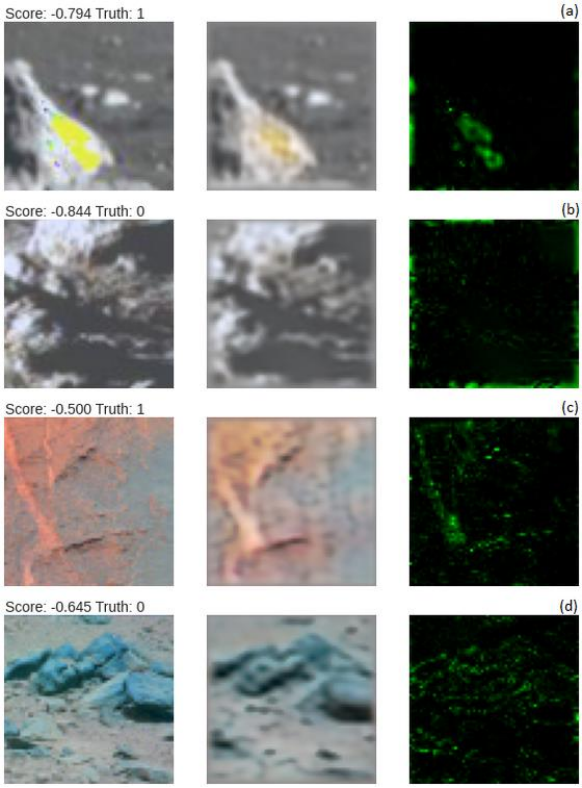

Figure 1: Representative images of the input (left), reconstruction (middle), and squared-difference error of the green channel (right) for (a) novel and (b) typical Lunar images, and (c) novel and (d) typical Martian images.

Our goal is to develop an end-to-end novelty detector that can be used operationally in planetary exploration missions. As a first step toward this, we have developed a novelty detection processing and evaluation pipeline for the purpose of algorithmic testing. In these preliminary stages, focus was placed on building a pipeline that supports rapid model prototyping, parameter optimization, and hypothesis validation. So far, the processing pipeline has been equipped with datasets from: The Moon and Mars.

In this paper, we introduce our processing pipeline, present structural improvements upon an existing baseline network and demonstrate its capabilities. We also implement a variational autoencoder (VAE) and gauge its performance against this baseline model. We explore the predictive capabilities of autoencoder-based novelty detection algorithms, expanding the quantitative analysis to the viability of such networks in an operational setting. In many ways, this study represents the strides that have been made to date towards an end-to-end detector that can be integrated into future rover missions on the systems level.

2 RELATED WORK

Novelty detection is part of a well-established set of techniques used to detect samples or features from within a set of data that are either unique or statistically uncommon. In nomenclature, *anomaly detection* is an umbrella term that covers both outlier detection and novelty detection. Although the definitions of these methods differ slightly, it is often appropriate to use them interchangeably [2].

Autoencoders have been applied to the problem of novelty detection since the early 2000s [3]. However, interest in the approach picked up in the mid 2010s when several teams released research demonstrating the applicability of autoencoders to image-based novelty detection [4], [5]. Extensions to the standard fully convolutional autoencoder were proposed that use *probabilistic* encoders and decoders—that is, they output parameters of the encoding distribution instead of the encoded pixel values themselves. Two primary extensions have been proposed [6], [7], [8]:

1) ***Variational autoencoders (VAEs)*** leverage the Evidence Lower Bound and KL-divergence to map the latent space representation to a prior distribution, typically a unit Gaussian [9], [10]. The latent representation is found by sampling from the encoded distribution, while the final data product is obtained by decoding the latent representation back to the original dimensionality.
2) ***Adversarial autoencoders (AAEs)*** leverage an adversarial procedure to obtain reconstructions. They calibrate the aggregated posterior of the latent distribution by matching it to an imposed prior distribution [11].

Novelty detection as applied to the domain of planetary exploration was spearheaded by Kerner et al. [1], [12]. In these works, the authors established a dataset of the Martian terrain for the purpose of developing and testing novelty detection algorithms. They implemented and analyzed a swath of techniques, including PCA, Reed-Xiaoli detectors, GANs, and CAEs. Various loss functions and novelty scores were used to compare the advantages and disadvantages of each detector. It was determined that, while autoencoders only performed on par or marginally better than other methods, they were easier to visualize and thus could be used to added more meaningful context to detections than alternative approaches.

3 METHODOLOGY

In this work we present a novelty detection processing and evaluation pipeline that leverages TensorFlow 2 to design and build autoencoder networks. As the purpose of this pipeline is to conduct rapid model development, many network configurations have been tested to date. Here, to represent the processing pipeline's capabilities, we introduce two preliminary models created using the pipeline. The first model is a baseline CAE, the other is a VAE. Although adversarial methods show promise in the area of novelty detection, their analysis has been saved for future research.

So far, the processing pipeline has been equipped with the Martian dataset introduced in [1], and with a preliminary dataset of the Lunar environment. Since the Martian dataset was already established in the field, it served as a sound starting point to begin building and testing architectural variants. It also served as a platform to automate the training procedure and build a code base that was highly modular, allowing quick parameter tuning, training, and performance metering. Once a reasonable baseline model was built using the Martian data, we optimized the baseline on the Martian data, expanded our research to assess the utility of a VAE on the same data, and retrained the baseline CAE on the Lunar set.

3.1 Datasets

Martian Dataset. The Martian dataset, derived from Kerner et al.'s work in [1], is composed of images taken by the Curiosity rover using its Mastcam imaging system. The resulting 6-channel multi-spectral images span the 400-1100nm range [13]. Novelties were labelled on the image level, and were defined to be either geologic anomalies (meteorites, floating rocks, veins) or the direct result rover actions (drill hole tailings or wheel scuffed rocks); all other images were labelled typical (Fig. 1(c,d)). Each image was cropped such that input images had dimension 64x64x6. Only typical images were used for training (9302 images) and validation (1386 images). When testing, roughly half of the data was typical (426 images), the other half was novel (430 images).

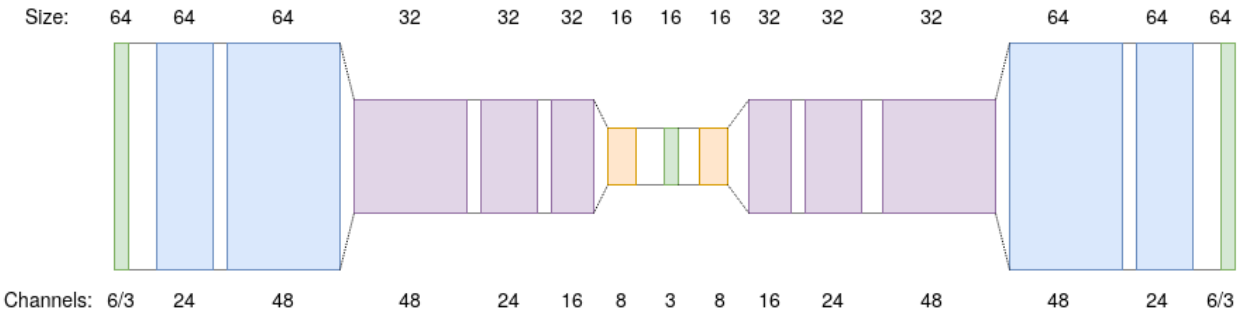

Figure 2: Architecture of the baseline convolutional autoencoder (CAE). Input channels vary between 6 for the Martian set, and 3 for the Lunar set.

Lunar Dataset. The Lunar dataset is composed of images taken from the Chang'E-4 Lander's Terrain Camera (TCAM) and Yutu-2's Panoramic Camera (PCAM) [6]. Images from both the TCAM and PCAM are 3-channel RGB images spanning the 420-700nm range. Novelties in these datasets were defined to be suspected meteorites, young craters, and image artifacts on otherwise normal surface features, such as saturated pixels. In total, the Lunar set contained 278 images. To maintain consistency with the Martian set, each image was fragmented into 64x64 patches, making the input dimension 64x64x3. In this study only the PCAM images were used for training and testing. Though the TCAM images were available and labelled, computational considerations early the pipeline's life prohibited use of the whole set.

After tiling the 168 PCAM images into 64x64 patches, 90,720 images were created with 39 novelties. Filtering was carried out to remove image patches from portions above the horizon whose pixels were all, or nearly all black. Other artifacts, such as green channel dominance and discolorations were difficult to filter and hence were kept in the Lunar set. In the end the training set consisted of 69,840 images, 12.5% of which were set aside for validation. The test set contained 10,940 typical images and 39 novel images.

Dataset Generation. We used a GTX 1080 Ti graphical processing unit (GPU) to accelerate our training cycles; with almost 70'000 images in the Lunar set, we opted to use generating functions to read our data as opposed to loading them into memory all at once. Since the processing pipeline is expected to be used for algorithmic experimentation on large planetary datasets, and with the eventual objective of running the algorithm in operational field tests, this decision fits with our project goals.

During training, data is pre-fetched into the computer's memory buffer. Immediately before being loaded into the GPU for training, the images are batched, mapped through a normalizing function, and cached for subsequent epochs. As described above, some image filtering was conducted for both the Martian and Lunar sets; these filtering operations are implemented in the generating function itself. In terms of speed, this approach prolonged the data reading process somewhat (as opposed to pre-filtering the data). However, it provided increased flexibility when managing multiple datasets, invariably with different formatting, structure, and required augmentations.

3.2 Baseline Convolutional Autoencoder

The baseline convolutional autoencoder has the structure shown in Fig. 2. The dimension of the latent space is 16x16x3 (green, middle), which marks a 32-times reduction for the Martian dataset, and a 16-times reduction for the Lunar dataset. Various encoder/decoder depths were tested; in most cases, deeper networks tended to increase performance. The baseline CAE in Fig. 2 has six convolutional layers and six transpose convolutional layers in the encoder and decoder respectively. Each reduction in image size was achieved using a double stride. Similarly, in the decoder, to increase the image size a stride of two was used during transpose convolutional operations.

Preliminary experiments indicated that the network provided in [1] could be improved by increasing the capacity of the model. Our intention was that, by increasing the capacity of the model, more subtle features could be extracted from both the encoding and decoding steps, boosting detection performance in structurally subtle images.

Normalization. Normalization for the data offered an unforeseen avenue of exploration when developing the baseline CAE. Work done in [1] used a normalization approach whereby the images were normalized along the batch dimension only. With this method, normalization statistics were calculated pixel-wise, each pixel

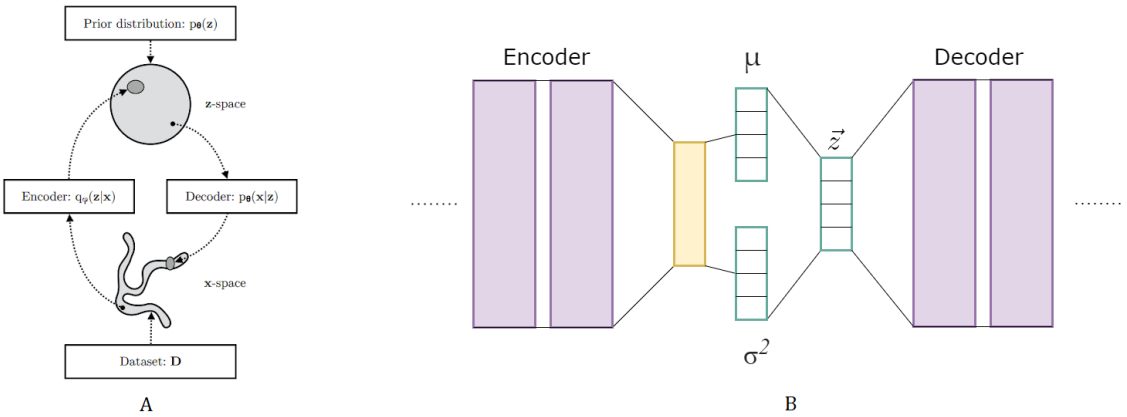

Figure 3: (a) Visualization of a CVAE. (b) Diagram of encoder and decoder interface; the output of the encoder is the Gaussian approximation of the posterior distribution.

location being standardized to a unit Gaussian. We implemented this normalization tactic in our pipeline, naming it 'standard pixel-wise'. We found max-min scaling methods particularly useful for visualizing the images, however, convergence was less effective. The normalization method that was eventually selected is like the standard pixel-wise approach, but it calculates the statistics over each channel as well. This way, each channel is mapped to a unit Gaussian independent from the other channels. This approach has the advantage that channel-wise outliers are mapped further from the mean, conceptually allowing them to be easily identified with a variance threshold.

Parameter Selection. Batch normalization was built into the baseline CAE and immediately improved convergence when training. Learning rates were varied between 0.00001 and 0.1. It was found that tuning the learning rate also improved network convergence considerably, especially after adding or removing network capacity. The learning rate that was selected for the baseline CAE was 0.001. A leaky ReLU activation was used with a rate of 0.1, the kernel size for all convolutional operations was set to 5, zero padding was employed in all layers., and batch size was set to 100. For regularization, early stopping was implemented with a patience of 10 epochs and dropout layers were added with a mild drop-rate of 0.05. Managing the patience and dropout rate was troublesome in the early stages. When the patience was below 3 epochs and a drop-rate higher than 0.1 was used, the network would stop training prematurely. The main reason for this is likely because when the drop-rate was high, training and validation losses became volatile, particularly when the learning rate was also high. This volatility would lead to validation loss increases for several epochs, at which point the training would be halted. When the number of trainable parameters was increased, the patience was set to 10, and the learning rate was set correctly, this effect was diminished.

Loss Functions. The loss functions used include mean squared error (MSE), structural similarity index (SSIM), and a linear combination of the two functions (Hybrid). The relative contribution of the MSE and SSIM terms are controlled by a tuning parameter λ. The λ parameter played an important role in optimizing the training procedure for Hybrid-trained networks. Though many λ parameters were tested, the baseline CAE presented here used a value of λ=0.1. An in-depth description of each of these loss functions is given in [1].

Novelty Scores. The novelty scores were calculated between the original input image and the reconstructed output. Implemented scores include (i) L2 norm (ii) Outlier Count, a score proposed in [1] to assess how many pixels fall above the mean pixel intensity, and (iii) the loss function itself.

Performance Metrics. Since novelty detection is effectively a binary classification problem, many metrics have already been developed to monitor performance. One of the most well known is the receiver operating characteristic (ROC) curve, which plots the true positive rate as a function of the false positive rate. Other metrics used include Precision @ N, and precision-recall curves. All of these metrics have been built into the evaluation portion of the processing pipeline. Other metrics that are less common but have been found to provide insight into the performance of the detector were also included. One such example is an F1-score versus threshold plot. Since we are interested in building a detector for systems level operations, it was essential to apply metrics more tangible than general performance curves. The most conventional of these is the confusion matrix of the classifier; although it is difficult to compare general performance *between* detectors with confusion matrices, it is easier to understand the performance of a single detector.

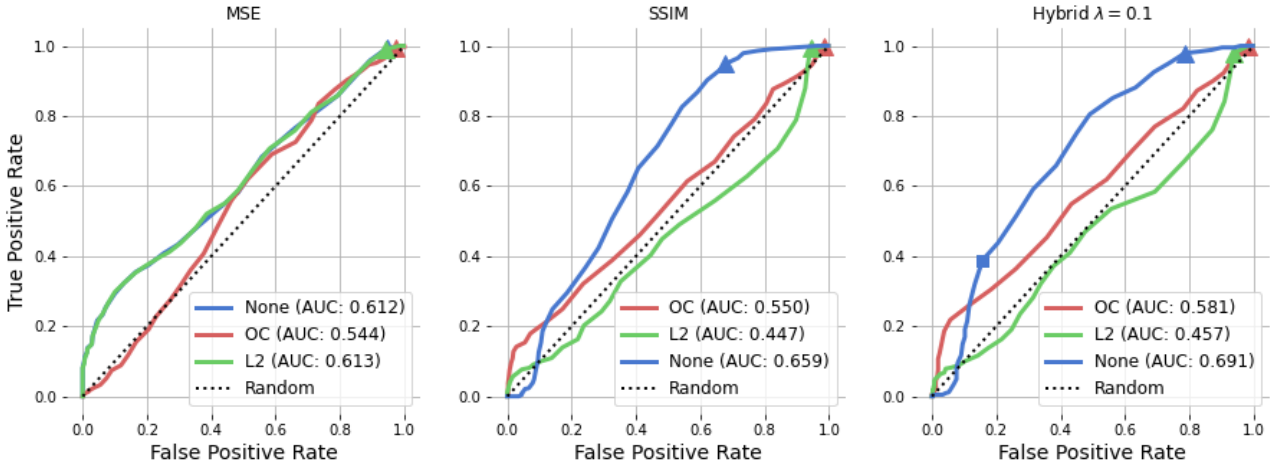

Figure 4: ROC curves for baseline CAE when trained on the Martian set with (a) MSE, (b) SSIM, and (c) Hybrid losses.

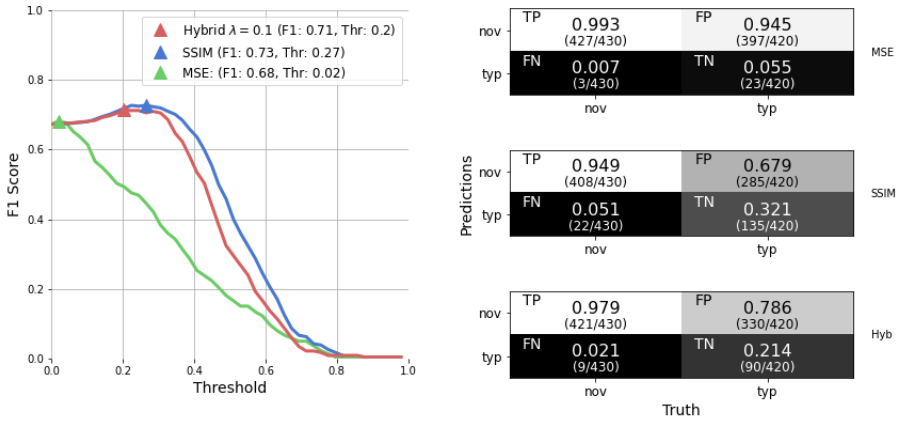

Figure 5: (a) F1 score as a function of normalized threshold for different loss functions and (b) the confusion matrices at the best F1 score.

3.3 Variational Autoencoder

The main difference between a Variational Autoencoder (VAE) and a CAE is that instead of directly mapping the input image into the latent space, the VAE attempts to model the distribution of the latent space (Fig. 3).

A Convolutional Variational Autoencoder (CVAE) parameterizes the posterior distribution with a Convolutional Neural Network (CNN). Outputting the means and variances of a multivariate Gaussian in the latent space, a CNN based decoder then consumes the sample generated from the posterior and reconstructs the image to its input dimension.

Loss Functions. The algorithm aims to maximize the Evidence Lower Bound (ELBO) by optimizing the Kullback-Leibler divergence between the approximated posterior and a standard Gaussian (the imposed prior). This forces the latent distribution to be as close to a standard Gaussian as possible.

Novelty Scores. The CVAE was trained only on typical images such that it generated an approximation of the distribution of the latent space particular to typical images. Since the encoder approximates the latent multivariate Gaussian, a natural choice for a novelty score is the distance between a random sample from the learned latent distribution on a test data point and the approximated distribution of the latent space. In this case, the Mahalanobis Distance [14] will be used as the distance metric. The intuition is that when typical images are encoded the predicted distribution parameters will resemble the distribution approximated by the encoder during training, and hence have a low novelty score.

4 RESULTS

Parameter selection for the baseline CAE was

conducted on the Martian dataset first. After the baseline CAE reached an acceptable level of performance, it served as a benchmark when applying it to the Lunar terrain and when implementing the VAE.

4.1 Baseline Convolutional Autoencoder: Mars

ROC curves for the baseline CAE are shown in Fig. 4. Here it can be seen that the best performance was achieved with the Hybrid loss at an Area Under the Curve (AUC) of 0.691. This outperforms the best detector from [1], also using a Hybrid loss, with an AUC of 0.650. Performance improvement for the SSIM-trained detector was also observed when the loss function itself was used as the novelty score. These improvements come at the cost of performance volatility. For example, when scoring images with the L2 norm for either SSIM or Hybrid-trained models, performance was worse than a random classifier. In contrast, the network varieties given in [1] do not fall below AUC=0.56, and thus have a lower AUC variance across detection methods. For each loss employed during training, the best results were obtained when scoring with the loss itself. Across the board, the Outlier Count score performed only marginally better than random.

Triangles in Fig. 4 denote the location of the highest F1 score, which measures the classifier quality using both the precision and recall. The true positive rate at each of the best F1 scores stays above 0.9; however, the corresponding false positive rates remain high, above 0.6 in all cases. The profile of the F1 score versus the (normalized) threshold is shown in Fig. 5 for each loss tested. Here, the MSE curve has its best F1 score at a threshold of 0.02; similar thresholds were found for all poor performance classifiers. Low thresholds on this level indicate that the best F1 score is found when the detector is heavily biased towards novel class predictions.

Notice that the two highest performing detectors also had the two highest thresholds aligned with their best F1 score. On the right side of Fig. 5 the confusion matrices at the best F1 scores are shown, where the columns are normalized to the number of actual class samples. As expected from the curves in Fig. 4, even at thresholds of 0.27 and 0.20, for the SSIM and Hybrid losses respectively, the majority of the total samples are predicted to belong to the positive (novel) class.

Although the Hybrid-trained detector has the highest AUC, the SSIM-trained detector has a slightly higher best F1 score. As shown in Fig. 5, this difference is exhibited by a lower false positive rate and a higher true negative rate; true positive and false negative rates between the SSIM and Hybrid detectors are consistent.

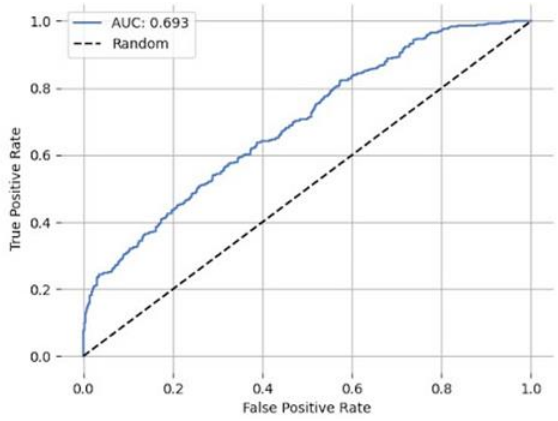

Figure 6: ROC curve for the CVAE novelty detector.

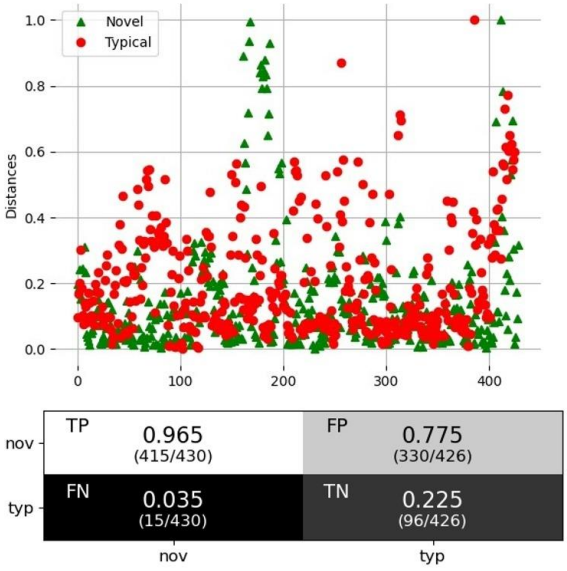

Figure 7: (top) Distances between sampled data points and individual test points and (bottom) confusion matrix at best F1 score for CVAE. Best F1 score: 0.706 at a threshold of 0.025.

4.2 Variational Autoencoder: Mars

Displayed in Fig. 6 is the ROC curve for the CVAE novelty detector. It achieved an AUC of 0.693, outperforming the result from [1] and performing slightly better than the Hybrid loss for the baseline CAE.

Looking at the confusion matrix in Fig. 7(bottom), at the threshold that outputs the best F1 score, the classifier tends to catch mostly novel images, with a recall value of 0.965. However, the CVAE tends to misclassify most of the typical images as novel, achieving a precision of only 0.555. This combined with the low

threshold indicates that the classifier biases its predictions towards novel class labels. Supplementary to this, Fig. 7(top) shows the Mahalanobis Distance of each data point in the test set. Separation is observed for novel data points at distances above 0.6. However, at distances below 0.4 novel and typical data points are clumped together, highlighting a limitation of distance metrics for novelty detection.

4.3 Baseline Convolutional Autoencoder: Moon

Resulting AUCs are displayed in Tab. 1 when applying the baseline CAE with the same parameters defined in Section 3.2 to the Lunar set. Tab. 1 also contains Hybrid-trained results when the λ parameter was tuned between 0.01 and 1.0. The ROC curves of the four best performing detectors is shown in Fig. 8. As was the case on the Martian set, the Hybrid-trained and SSIM-trained networks outperformed their MSE variants. One marked difference between the Martian and Lunar set is the efficacy of the L2 norm score for networks trained with SSIM and Hybrid losses. Although the L2 norm was the worst performing score on the Martian set, it consistently gave the highest AUC for each of the Hybrid networks, independent of the tuning parameter (Tab. 1).

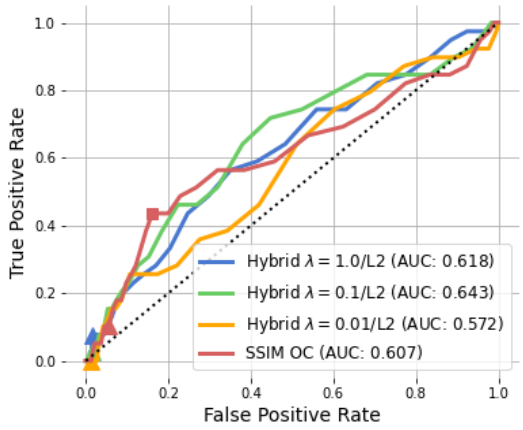

Figure 8: ROC curves baseline CAE detectors trained on the Lunar set with Hybrid and SSIM losses.

5 DISCUSSION

The results presented demonstrate noticeable improvements to current state-of-the-art novelty detectors. When trained on the Martian set, the baseline CAE outperformed the detector proposed in [1] by 6.3%, while the VAE outperformed by 6.6%.

5.1 Operational Viability

Operational viability refers to the ability of an algorithm to realistically help guide mission-level objectives. To understand this aspect more concretely for the networks described in this study, it helps to consider specific quantitative results. The profile of an ROC curve provides this type of tangible insight. For instance, one point along the ROC profile that has already been used to assess the performance of a single detector is the threshold that yields the highest F1 score (Fig. 5). Unfortunately, the false positive rates were far too high to return reliable detections. Due to the class imbalance, the high false positive rates that were observed on the Martian set become dramatically more pronounced on the Lunar set.

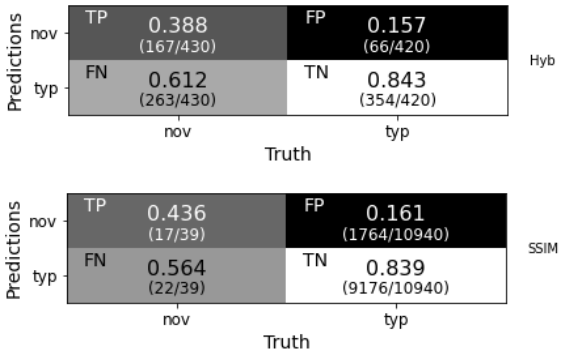

Figure 9: Confusion matrices of the highest true positive to false positive ratio on (top) the Martian set and (bottom) the Lunar set.

Table 1: Area under the ROC curve for different losses and novelty scores when the baseline CAE was trained on the Lunar set.

Loss	Score	AUC
MSE	L2 norm	*0.494*
	Outlier Count	*0.300*
	Loss	*0.499*
SSIM	L2 norm	0.605
	Outlier Count	**0.607**
	Loss	*0.442*
Hybrid $\lambda = 1.0$	L2 norm	**0.618**
	Outlier Count	*0.484*
	Loss	0.548
Hybrid $\lambda = 0.1$	L2 norm	**0.643**
	Outlier Count	0.534
	Loss	0.506
Hybrid $\lambda = 0.01$	L2 norm	0.572
	Outlier Count	0.505
	Loss	0.406

To see this in action, the confusion matrix of another point has been selected, denoted by squares in Fig. 4(c) and Fig. 8. These points were chosen because they represent the location of the highest true positive to false positive ratio. The associated confusion matrices are

shown in Fig. 9. While the false positive rates only differ by 0.004, the class imbalance on the Lunar set invokes approximately *100-times less* true positives than false positives, as opposed to *2.5-times more* on the Martian set. It is worth noting that the imbalanced test set in the Lunar dataset is more representative of operational conditions; in reality, novel data will by definition make up a small fraction of considered images. One way to consider the utility of the novelty detector is as a first pass before sending images labeled 'novel' on for human inspection. In this role, it has boosted the proportion of novel images from 0.36% (39 out of 10,979) to 0.95% (17 out of 1,781).

5.2 Future Work

The work presented in this paper leaves ample room for future research, some of which is already underway. Perhaps the most critical development needed to fully address the operational viability of a Lunar novelty detector is a more established dataset. Currently, in partnership with Mission Control Space Services, we are constructing an indoor Lunar analogue terrain site. Beyond the sheer number of images, more consistent and deliberate labels are crucial, especially for the purpose of novelty detection where contamination of novel samples in the training set has been shown to decrease novelty detection performance [10].

Another potentially fruitful line of research is presented in [12]. Here, the authors leveraged a CAE to create reconstruction error maps that were then fed to a convolutional neural network (CNN) for binary classification. A benefit of this approach is that the CNN can extract novel features from within the images themselves. This contrasts the methods used in this study where classifications were conducted by setting thresholds on global image properties. Preliminary research in this direction have shown positive results.

6 CONCLUSIONS

In this work, a processing pipeline was introduced that emphasizes fast model development and performance monitoring for the task of novelty detection. Results from tests on two autoencoder variants were shown and experiments were run on a preliminary Lunar dataset. The baseline CAE, based on the work done in [1], incorporates multiple design changes and was shown to achieve high performance under certain conditions. However, this came at the cost of stability across detection methods. The VAE performed at the same level as the baseline CAE on the Martian set, and though some separability was seen between novel and typical samples, it was not enough for distance-based metrics to be fully effective.

Operational viability considerations were highlighted for specific threshold selections of the baseline CAE. Though both the CAEs and VAE were biased towards predicting the novel class when using the best F1 score, when more intentional threshold selections were made according to the ROC profile, operational viability improved. Regardless, the detectors are currently insufficient to be used operationally. They may move deeper into the viable regime by dramatically lowering the false positive rate such that the number of truly novel predictions dominates the total number of positive predictions. Future work involves the development of an established, well labelled Lunar dataset, the application of more complex autoencoding frameworks, and the addition of a CNN for binary classification.

Acknowledgements

We would like to acknowledge Mission Control Space Services for their collaboration and funding support.

References

[1] Kerner, H.R., Wagstaff, K.L., Bue, B.D. *et al.* (2020). Comparison of novelty detection methods for multispectral images in rover-based planetary exploration missions. *Data Min Knowl Disc*.
[2] Xu, X. Liu, H. Yao, M. (2019) Recent Progress of Anomaly Detection. *Complexity*, vol. 2019.
[3] Thompson, B. B. *et al.* (2002*).* Implicit learning in autoencoder novelty assessment, *Proc. of the 2002 IJCNN*. vol. 3 pp. 2878-2883.
[4] Chen, J. Sathe, S. Aggarwal, C. and Turaga, D. (2017). Outlier detection with autoencoder ensembles, *Proc. 17th SIAM Int. Conf. Data Mining*. pp. 90–98.
[5] Pimentel, M. A. F. Clifton, D. A. Clifton, L. and Tarassenko, L. (2014). A review of novelty detection. *Signal Processing*, vol. 99, pp. 215–249.
[6] An, J. and Cho, S. (2015). Variational Autoencoder based Anomaly Detection using Reconstruction Probability. *SNU Data Min. Cent*. Special Lecture.
[7] Leveau, V. Joly, A. (2017) Adversarial autoencoders for novelty detection. *Workshop track – ICLR.*
[8] Beggel, L. Pfeiffer, M. and Bischl, B. (2019) Robust Anomaly Detection in Images using Adversarial Autoencoders.
[9] Doersch, C. (2016) Tutorial on Variational Autoencoders. pp. 1–23.
[10] Yang, X. (2017) Understanding the Variational Lower Bound. pp. 1–4.
[11] Makhzani, A. Shlens, J. Jaitly, N. Goodfellow, I. and Frey, B. (2015) Adversarial Autoencoders.
[12] Kerner, H. R. Wellington, D. F Wagstaff, K. L. *et al.* (2018). Novelty detection for multispectral images with application to planetary exploration. *Neural Inf. Process. Syst*.
[13] Bell, J. F. et al. (2017). The Mars Science Laboratory *Curiosity* rover Mastcam instruments: Preflight and in-flight calibration, validation, and data archiving. *Earth and Space Science*, 4, 396–452.
[14] Mahalanobis, P.C. (1936) On the generalized distance in statistics. *Proc. Natn. Inst. Sci. India*, 2, 49–55.